Towards standardized 3DTV QoE assessment: Cross-lab study on display technology and viewing environment parameters

Marcus Barkowsky[a], Jing Li[a], Taehwan Han[d], Sungwook Youn[d], Jiheon Ok[d], Chulhee Lee[d], Christer Hedberg[b], Indirajith Vijai Ananth[b], Kun Wang[bc], Kjell Brunnström[bc], Patrick Le Callet[a]
[a]LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, France; Email: Firstname.Lastname@univ-nantes.fr; [b]Dept. of Netlab, Acreo Swedish ICT AB, Sweden; Email: Firstname.Lastname@acreo.se; [c]Mid Sweden University, Sweden; [d]School of Electrical and Electronic Engineering, Yonsei University, South Korea; Email: Lastname@yonsei.ac.kr

ABSTRACT

Subjective assessment of Quality of Experience in stereoscopic 3D requires new guidelines for the environmental setup as existing standards such as ITU-R BT.500 may no longer be appropriate. A first step is to perform cross-lab experiments in different viewing conditions on the same video sequences. Three international labs performed Absolute Category Rating studies on a freely available video database containing degradations that are mainly related to video quality degradations. Different conditions have been used in the labs: Passive polarized displays, active shutter displays, differences in viewing distance, the number of parallel viewers, and the voting device. Implicit variations were introduced due to the three different languages in Sweden, South Korea, and France. Although the obtained Mean Opinion Scores are comparable, slight differences occur in function of the video degradations and the viewing distance. An analysis on the statistical differences obtained between the MOS of the video sequences revealed that obtaining an equivalent number of differences may require more observers in some viewing conditions. It was also seen that the alignment of the meaning of the attributes used in Absolute Category Rating in different languages may be beneficial. Statistical analysis was performed showing influence of the viewing distance on votes and MOS results.

Keywords: Subjective assessment, Viewing environment, Stereoscopic Displays, 3D Quality of Experience, Cross-Lab Validation, Standardization

1. INTRODUCTION

Reliable and reproducible subjective measurement of Quality of Experience (QoE) in 3DTV is currently investigated for optimizing 3D service parameters and as a necessary prerequisite towards the development of objective models. QoE for 3DTV is known to extend over several psychophysical dimensions such as picture quality, depth sensation, and visual comfort which may be combined to higher level indications such as naturalness, presence and visual experience [1].

The perception of degradations measured in subjective assessment studies is influenced by the viewing conditions. In stereoscopic 3DTV, selecting and calibrating the display may be more important than in 2D, as additional technological factors for the display such as maximum perceived brightness or crosstalk may have significant influence, and may be difficult to measure across subjective assessment labs [2][3].

The influence of the viewing environment, such as illumination, viewing distance, voting interface, observer screening, training and introduction to the experiment, is expected to differ significantly from the influence that was perceived when 2D recommendations, such as ITU-R BT.500, were established. For example, in line-interleaved passive, polarized displays, the horizontal resolution often exceeds the vertical resolution per view by a factor of two, leading to questions regarding the appropriate viewing distance of 3H or 5H for Full-HD content.

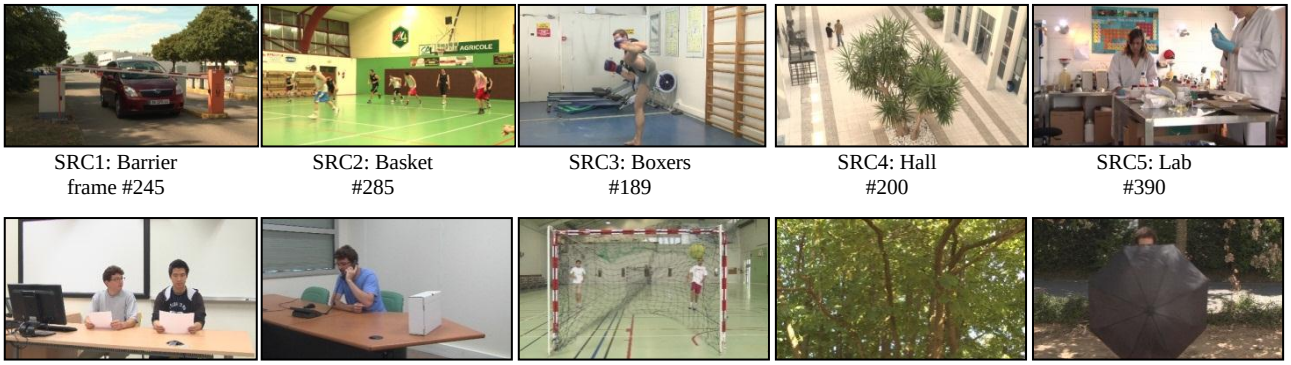
Recently, the Video Quality Experts Group (VQEG) started a re-evaluation of the viewing conditions and the display specifications in preparation of new recommendations in standardization organizations such as ITU and EBU [4]. A freely available, common set of video sequences has been published that contains only symmetric video coding and resolution reduction as degradations. Measurement only on the video quality scale may therefore be sufficient, as opposed to asymmetric video coding, changes in camera distance, or transmission errors which are related to visual discomfort and depth realism [5][6].

This paper analyzes the results obtained from three subjective experiments on the aforementioned database. The experiments have been done in three different locations, with different equipments and viewing conditions: At Acreo Swedish ICT AB in Sweden, at the IRCCyN lab of the University of Nantes in France, and at the Yonsei University in South Korea. Section 2 introduces the properties of the evaluated video sequences. The subjective experiment design is explained in Section 3, and the obtained observer votes are analyzed in Section 4, followed by the conclusion in Section 5.

2. VIDEO CONTENT AND DEGRADATIONS

A detailed description of the video content and the applied distortions has been published in [7]. A short summary for the Nantes-Madrid 3D Stereoscopic source sequences (NAMA3DS1) and the distortions introduced in the Coding and Spatial Degradations (COSPAD1) dataset [8] is presented in the following.

The source content (SRC) has been selected in order to provide a wide variety of different content types as can be seen in Figure 1. All sequences have been captured using a Panasonic AG-3DA1E twin-lens camera at 1920x1080p25 resolution. Most of them have durations of 16 seconds and were stored uncompressed on a ClearView Extreme System. SRC2, SRC3, and SRC5 have been stored on the SD cards of the camera at a maximum bitrate of 24Mbit/s per view, and SRC10 has a length of 13 seconds.



SRC1: Barrier frame #245 | SRC2: Basket #285 | SRC3: Boxers #189 | SRC4: Hall #200 | SRC5: Lab #390

SRC6: News report #150 | SRC7: Phone call #181 | SRC8: Soccer #193 | SRC9: Tree branches #200 | SRC10: Umbrella #230

Figure 1: Source sequence thumbnails

The degradations, called Hypothetical Reference Circuits (HRC), are summarized in Table 1. They have been chosen to exhibit mostly perceptual impairments on the image quality scale, i.e. avoiding influences on depth realism or visual comfort. In some cases, this may not hold true, in particular, the strong coding degradations in HRC3 and HRC4 may lead to a sensation of binocular rivalry and therefore lead to visual discomfort. Depending on the content characteristics, in high frequency regions spatial details may be lost, therefore also loosing exploitable disparity information, leading to less perceived depth effect. For example, in SRC9 the depth differences between the leaves may no longer be perceived.

Table 1: Degradations

HRC	Impairments and Degradations	
	Type	**Parameters**
0	None – Reference sequence	
1	Video coding (H.264)	QP 32
2	Video coding (H.264)	QP 38
3	Video coding (H.264)	QP 44
4	Still image coding (JPEG2k)	2 Mb/s
5	Still image coding (JPEG2k)	8 Mb/s
6	Still image coding (JPEG2k)	16 Mb/s
7	Still image coding (JPEG2k)	32 Mb/s
8	Reduction of resolution	↓4 downsampling
9	Image sharpening	Edge enhancement
10	Downsampling & sharpening	HRC 8 + HRC 9

3. SUBJECTIVE EXPERIMENT SETUP

3.1 Viewing environment and displays

In total, four different conditions were used by the three laboratories in Sweden, South Korea and France as shown in Table 2. In all experiments, the maximum display brightness perceived through the polarized or active shutter glasses was measured and the background illumination was adjusted to 15% as specified by ITU-R BT.500. The main differences were in terms of language, display technology, and number of observers as well as the voting device. At Acreo half of the sequences were presented at a viewing distance of 3H, the other half was shown at 5H. Additional observer information was obtained in the labs.

Table 2: Viewing environment

Experiment	EXP1	EXP2	EXP3a	EXP3b
Laboratory	IRCCyN, Nantes	Yonsei, South Korea	Acreo, Sweden	
Display	Philips 46PFL9705H	Hyundai S465D	Hyundai S465D	
Technology	Active Shutter glasses	Polarized FPR, glasses	Polarized Frame-Pattern-Retarder (FPR)	
Viewing distance	3H (1.72m)	3H (1.72m)	3H (2.5m)	5H (4.2m)
Voting device	Screen	Paper	Screen	
Language	French	Korean	Swedish/English	
Number of observers	29	28	24	24
Obs. per viewing	1	2	1	1
Observer screening method	acuity, stereo-acuity, color	stereo-acuity, color	acuity, stereo-acuity, color	
Additional observer information	age, gender, 3D viewing experience, directing	eye distance	age, gender, 3D viewing experience, simulator sickness questionnaire	

3.2 Subjective Test Setup

Absolute Category Rating with Hidden Reference (ACR-HR) as specified in ITU-T P.910 was conducted in all labs. The training instructions were translated from a shared English version to the three native languages. At Acreo, half of the observers were native Swedish speakers, the other half performed the experiment in English.

All observers watched all 110 processed video sequences (PVS) in EXP1 and EXP2. In EXP3, the PVS were split in two groups. Each group contains all HRC of SRC2. The other SRC were equally distributed and the HRCs were selected in order to obtain a uniform distribution of Mean Opinion Scores (MOS) based on the results of the two other labs leading to the repartitioning as shown in Table 3. Half of the observers started viewing SetA at 3H and continued after a break with SetB at 5H, half of the observers started viewing SetB at 3H before watching SetA at 5H. The voting in EXP3 was performed on three scales simultaneously; “visual discomfort” and “sense of presence” will not be analyzed in this paper.

Table 3: Subset selection at Acreo

	HRC0	HRC1	HRC2	HRC3	HRC4	HRC5	HRC6	HRC7	HRC8	HRC9	HRC10
SRC1	A	B	B	B	A	A	A	B	A	A	B
SRC2	A,B	A,B	A,B	A,B	A,B	A,B	A,B	A,B	A,B	A,B	A,B
SRC3	B	B	A	A	B	B	A	A	A	A	B
SRC4	B	A	B	A	B	A	A	B	B	A	B
SRC5	A	A	A	B	A	A	B	A	B	B	B
SRC6	B	A	B	B	A	B	A	A	A	B	B
SRC7	A	B	B	A	A	A	B	A	B	B	A
SRC8	B	A	A	A	B	B	B	A	A	B	B
SRC9	A	A	A	A	A	B	A	B	B	B	B
SRC10	A	A	B	A	B	B	A	B	B	B	A

At the Yonsei University, EXP1, two observers were seated in front of the screen and the voting was written on paper, in the other labs only a single observer per session watched the PVS and a voting interface appeared on either a separate

screen (IRCCyN, EXP2) or on the same screen (Acreo, EXP3). EXP1 used an active shutter glasses system while the two other labs used the same polarized passive display.

4. RESULTS

4.1 Inter-lab Mean Opinion Score analysis

The Mean Opinion Scores (MOS) will be analyzed first across the different labs. For this analysis, the different viewing distances in EXP3 are not distinguished.

The Pearson Linear Correlation Coefficient (PC) and the Spearman Rank Order Coefficient (SROCC) as well as the scatterplots depicted in Figure 2 show that, in general, high correlation has been obtained between the different labs.

Slight differences can be seen in the usage of the voting scale, a linear regression curve has been fitted to the data, showing that the observers in EXP1 gave lower votes than those in EXP2, followed by those in EXP3, visible also in the grand mean values of 3.10, 3.13, and 3.37 MOS respectively. None of the differences is statistically significant on a 95% confidence using a student-t test or Wilcoxon rank sum test.

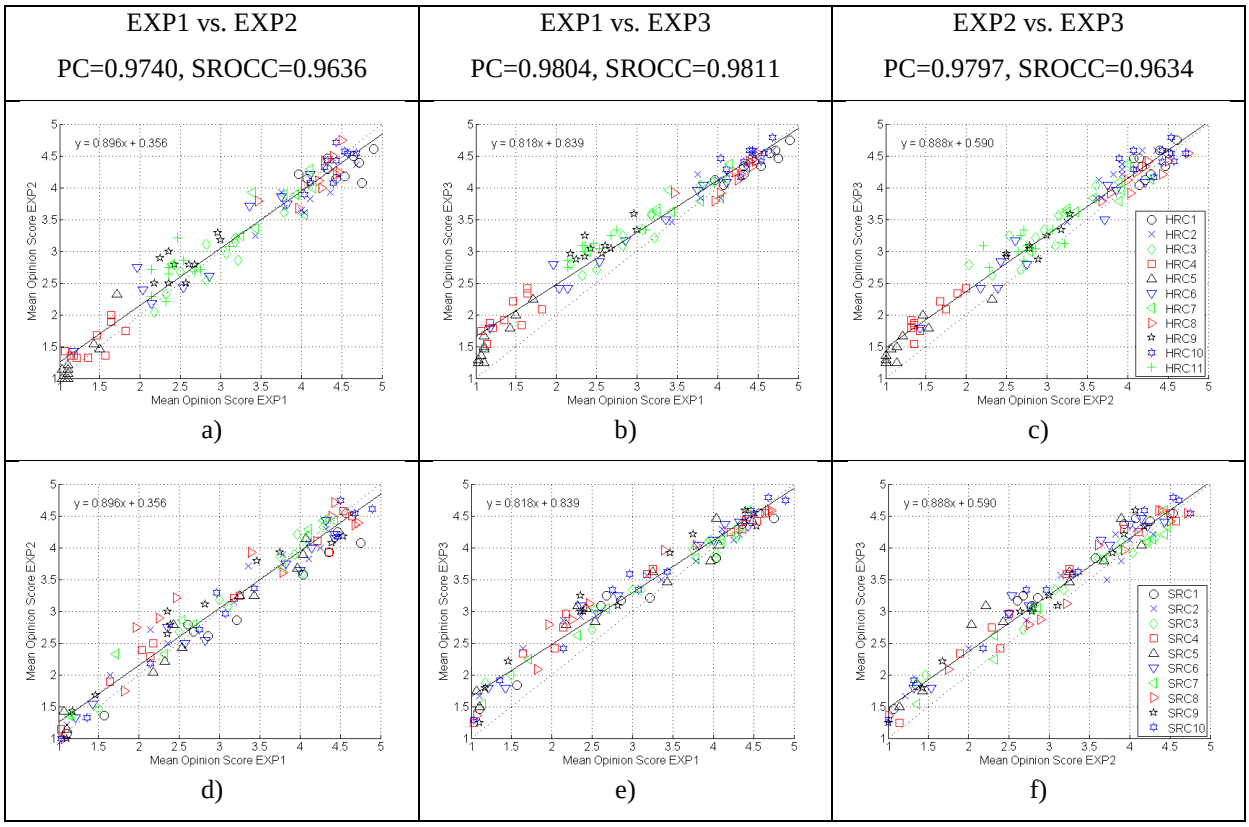


Figure 2: Inter-Lab scatterplots: a-c use different markers to distinguish HRCs, d-f to distinguish SRC

The largest divergence from the regression line is observed for comparing EXP1 and EXP2, the RMSE after fitting is 0.27, compared to 0.23 and 0.22 for (b) and (c) in Figure 2 which is in line with the rank order of the Pearson Correlation coefficients. The relation to the viewing environment conditions seems complex, no simple explanation can be provided at this moment.

Analyzing the influence of SRC sequences on the overall votes after linear fitting to EXP1, it can be found that the highest mean shift occurs for SRC1 (Barrier gate), which was voted with a MOS of 3.33, 3.02, and 3.16 in the three experiments. This may be due to different display technologies rendering the fine structures of the trees particularly interesting at the full resolution shutter glasses display. A display difference was expected for SRC2: The camera pan in this sequence has been criticized by several experts as leading to temporal sampling problems in shutter glasses displays.

The observer may get confused by alternating the left and right view while perceiving fast motion at the same time. Depth estimation and movement estimation coincide. However, this sequence does not show any particularity with MOS values of 3.12, 3.19, and 3.15.

The influence of HRCs on the overall voting shows the largest influence for HRC8, the downsampling of the resolution of a factor of 4. After alignment to EXP1, the three MOS values are 2.54, 2.75, and 2.77. While not being statistically significant, it may be seen that the observers perceive more degradation on the Full-HD active display in EXP1 than on the polarized screen which has half the vertical resolution per view in EXP2 and EXP3.

4.2 Number of significantly different PVS

In the analysis of subjective experiment studies, the conclusions often depend on obtaining statistical difference between PVS. In most cases, an increase in the number of subjects results in an increased number of statistically different PVSs. This has been analyzed for the 3 labs. The diagram in Figure 3 shows for a given number of observers the average number of statistical differences using a Monte Carlo simulation with 1000 trials. It can be observed that in EXP2, the same number of statistical differences can be obtained with approximately 3 more observers compared to EXP1. In EXP3 the number of additional observers is increased by approximately 7. Analyzing the source of this difference requires further subjective experiments.

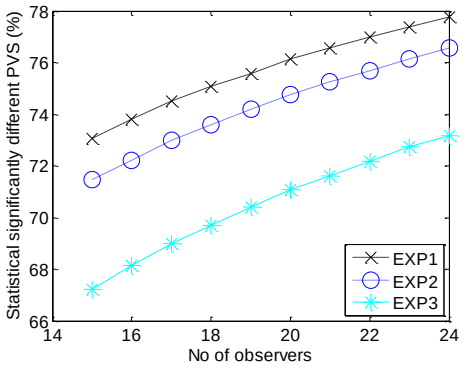


Figure 3: Statistically different PVS in per cent in function of number of observers

4.3 Influence of language on absolute category votes

Four different languages have been used in this subjective experiment: French, Korean, English, and Swedish. It is known that the Absolute Category Rating scale may be influenced by the meaning of the attributes in each language. Based on the research reviewed in [9], the individual votes were aligned to a common scale. Within the four languages, French was selected for the common scale because a complete experiment with 24 observers was available and mapping values to an absolute scale were published in the literature. First, the French votes were mapped to the values provided in [9]. Second, for each of the other experiments, the numerical vote values were mapped to the common scale using the assumption that the MOS values should coincide. The Root Mean Square Error (RMSE) was used as criterion. The five votes were therefore fitted on a minimum of 12 observers with 110 votes each for English and Swedish language in EXP3. Figure 4a shows the results of the alignment. It should be noted that the subjects taking the test in English were not native English speakers, which means that the difference on this language scale towards the others should be interpreted with caution. Most of the attribute rank orders correspond to the expected results from [9]. The remaining diagrams in Figure 4b show the scatterplots of the EXP1 as compared to the EXP2, and EXP3 with the corresponding languages used. The Pearson Correlation and Spearman Rank Order coefficients are integrated in the diagrams and show no significant improvement as compared to Figure 2.

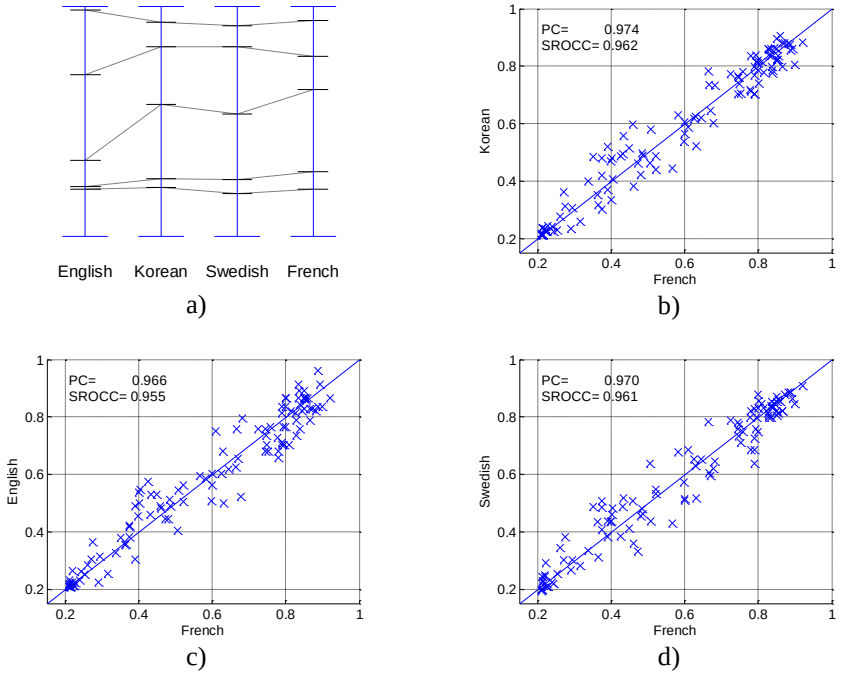


Figure 4: Alignment of adjectives in different languages

4.4 Comparing individual votes for viewing distances of 3H and 5H at Acreo

In EXP3, SRC2 was presented to all subjects both at 3H and at 5H viewing distance. Therefore, 24x11=264 pairs are available that compare the two viewing distances. A Two-way ANOVA analysis was performed with HRC and viewing distance as within-factors, showing statistical significance on both main factors, only slight significance for viewing distance ($F(1,23)=6.5$, $p=0.02$)[1], and strong significance for HRC as expected ($F(10,230)=73.73$, $p<0.01$). Analyzing the 11 conditions separately, it may be found that statistical difference is only present for HRC9, when the resolution is reduced by a factor of 4 and upscaled again ($M_{3H}=2.667$, $SD_{3H}=0.917$; $M_{5H}=3.173$, $SD_{5H}=0.761$, student-t $p=0.02384$, Wilcoxon $p=0.02854$).

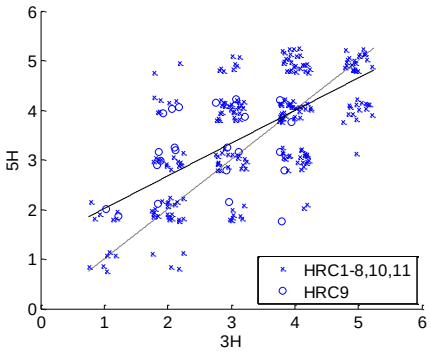


Figure 5: Scatterplot of observer votes for 3H and 5H for SRC2 of EXP3,
(Uniform random noise added to the absolute category rating votes from 1-5 for display purpose only)

[1] In the EXP3 the videos had been divided into two video sets containing one SRC (SRC 2) with all its HRC on both video sets as a common set. The common set is therefore not a pure within variable. Analyzing the viewing distance as a between effect on the common set gave, however, no significant difference. We believe though that it is more close to a within effect, since the videos of the common set were randomly mixed with the whole dataset.

Figure 5 shows a scatterplot of all votes, uniform random noise with amplitude of 0.5 was added to the quantized 5-point absolute category rating votes for displaying purpose only. HRC9 uses a different marker style, and the linear regression curve (solid line) shows a deviation from the main diagonal (dotted line) in favor of higher votes for 5H. The resolution reduction may be less perceivable at larger viewing distances, as 14 observers ranked them at least one attribute higher, while only 4 chose to prefer the quality in 3H over 5H; 6 observers were undecided. It should be noted that the analysis may be biased due to carry-over effects as all subjects started their viewing session in 3H.

The results indicate that observers may have seen a difference between 3H and 5H viewing distance individually but the influence on the Mean Opinion Score is limited. A larger experiment involving more observers and more PVS is required.

5. CONCLUSIONS

Quality of Experience assessment in stereoscopic 3D remains a challenging topic. The subjective experiment dataset used by the three subjective experiment labs presented in this paper was deliberately limited to the image quality degradation scale. The obtained Mean Opinion Scores were comparable although the lab setup differed within reasonable limits. Amongst the most important differences may be noted that active shutter glasses or passive polarized display technology was used, one or two observers judged the video sequence at the same time, the voting was performed either on paper, on the presentation display, or on a separate display. The analysis showed that differences may also be related to the meaning of the adjectives on the ACR voting scale in the different languages. Last but not least, the viewing distance on passive polarized screens was evaluated showing that observers tend to perceive less degradations when seated at 5H, corresponding to the vertical resolution per view, than at 3H, corresponding to the horizontal resolution per view.

Further experimental studies are required to verify the obtained results, including more observers and further viewing conditions. This research has been enabled by the availability of a freely available dataset and will be boosted by collecting the observer's votes in different viewing conditions, allowing further statistical analysis of the influence of the lab setup conditions.

REFERENCES

[1] Lambooij, M., IJsselsteijn, W., Bouwhuis, D. G., & Heynderickx, I., "Evaluation of Stereoscopic Images: Beyond 2D Quality," IEEE Transactions on Broadcasting, 57(2), 432–444, (2011).

[2] Tourancheau, S., Wang, K., Bulat, J., Cousseau, R., Janowski, L., Brunnström, K., et al., "Reproducibility of crosstalk measurements on active glasses 3D LCD displays based on temporal characterization," SPIE Electronic Imaging Stereoscopic Displays and Applications XXIII, 8288 (2012)

[3] Xing, L., You, J., Ebrahimi, T., & Perkis, A., "Assessment of Stereoscopic Crosstalk Perception," IEEE Transactions on Multimedia, 14(2), 326–337 (2012)

[4] Video Quality Experts Group (VQEG), "Test Plan on Evaluation and Specification of Viewing Conditions and Environmental Setup for 3D Video Quality Assessment," "http://www.its.bldrdoc.gov/vqeg/projects/3dtv/3dtv.aspx", (2012)

[5] Gutierrez, J., Perez, P., Jaureguizar, F., Cabrera, J., & Garcia, N., "Subjective assessment of the impact of transmission errors in 3DTV compared to HDTV," 3DTV Conference: The True Vision – Capture, Transmission and Display of 3D Video (3DTV-CON), (2011)

[6] Aflaki, P., Hannuksela, M. M., Häkkinen, J., Lindroos, P., & Gabbouj, M., "Subjective study on compressed asymmetric stereoscopic video," 17th IEEE International Conference on Image Processing (ICIP), pp. 4021–4024 (2010)

[7] Urvoy, M., Barkowsky, M., Gutiérrez, J., Cousseau, R., Koudota, Y., Ricordel, V., et al., "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," IEEE Fourth International Workshop on Quality of Multimedia Experience QoMEX, (2012)

[8] IRCCyN-IVC, "Nantes-Madrid 3D Stereoscopic database – Coding and Spatial Degradations NAMA3DS1-COSPAD1," "http://www.irccyn.ec-nantes.fr/spip.php?article1052", Nantes, (2012).

[9] Zielinski, S., Rumsey, F., & Bech, S., "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," Journal of the AES, 56(6), 427–451 (2008)